

Mining Urban Traffic Events and Anomalies

DM751

Abstract—As huge numbers of sensors relay real-time traffic measurements on city roads, real-time information management systems need to be in place that sieve relevant from irrelevant information. This is particularly important in mission critical applications. For example traffic operators need to be informed about potential anomalies caused by road traffic accidents or similar events in order to deploy timely mitigation strategies. We investigate a new method that elevates anomalies visually to an operator by using a combination of change-detection analytics and auxiliary information from human-sensors (e.g., through Twitter). We identify operational regimes of constant behaviour and change points that bound a regime. This allows us to account for drifts that are not captured in a model. This has the advantage over cumulative sum (CUSUM) algorithms that no threshold parameter needs to be specified upfront and relaxes the assumption of stationarity. Stationarity assumptions do not hold in traffic systems, as network effects take hold and traffic signal control intervenes when local traffic states are deteriorating.

I. INTRODUCTION

A. Background

In the city of Dublin, there are over two thousands inductance loop detectors and only a limited number of humans to monitor data from these detectors. These detectors count how many cars cross an intersection and provide a measure of traffic capacity. Measurements from these detectors are available to traffic operators in real-time in order to assist them in identifying potential problems in urban road traffic. However, the amount of data is overwhelming and operators face uncertainty and partial observability in incorporating this source of data into decision making processes. With the advent of Twitter, several authorities (e.g., police, smartphone navigation services) have started to broadcast useful information about urban traffic accidents and issues. It is a non-trivial task to verify all reported issues and assess whether any further action is needed to alleviate congestion.

Traditionally, public infrastructures, such as roads, rely on a small number of human operators monitoring numerous sources of real-time data. This approach is woefully inadequate and often too slow to mitigate traffic jams. Our goal is to apply data mining and machine learning methods to make sense of heterogeneous data from detectors and social media streams, and consequently improve traffic flow.

Various machine learning techniques already exist to tackle homogeneous sensor and social media data. We propose a heterogeneous approach to detect events, which gives more confidence in the output than homogeneous approaches. We quantify this improved confidence empirically in our Dublin traffic use-case.

In the case of homogeneous sensor data, change-detection [1], [2], clustering [3], and classification [4] methods are used to detect events in sensor data, and various state-estimation methods (e.g., [5]) are used to predict future traffic conditions. In particular, [6] applies estimation methods to predict traffic conditions on the same Dublin traffic data that we use. Non-stationary signals due to network effects, spontaneous changes, and drifts present a huge challenge to change detection methods.

User generated content can provide information outside the scope of traditional data sources [7]. For example, traffic sensors such as GPS units and loop detectors may identify delays or emerging traffic jams, but are unable to capture the whole picture in order to differentiate between an unplanned protest, unexpected weather conditions or a broken down car. Social media streams, such as twitter, have shown to be valuable information sources for such events [8], [7].

Social media provides a valuable source of real-time information and insights into the physical world. Starbird et al. examined how twitter users disseminated information and coordinated during a flood [7]. Sakaki et al. use a domain specific classifier to detect earthquake events [9]. Becker et al. use discovered clusters of related words or tweets to detect events [8]. Popescu et al. detect events from twitter using knowledge of known entities [10]. Nicholas et al. use twitter to automatically generate summaries of sporting events where key moments are detected through a rapid increase in the number of messages [11]. Much of this work addresses the challenges associated with event detection in the diverse and noisy twitter stream.

He et al. leverage social media content in order to improve the prediction of detector timeseries [12]. Specifically, they identify tweets with traffic indicators using an optimisation framework and were able to outperform auto-regressive moving averages models.

Daly et al. use sensor data to query Twitter in order to identify user generated messages that potentially ex-

plain current traffic conditions [13]. Our work presented in this paper performs the reverse situation and utilize traffic related tweets in order to probe sensor data as a validation to anomaly detection.

B. Our method

In this article, we are presenting an approach to detecting anomolous events from dedicated traffic detector measurements validated using social media (Twitter) data and statistical change-detection methods. We use natural language processing to extract location information from Twitter messages from authoritative sources and to identify explanations for anomalies. We employ a change-detection method for continuous-state Markov chains on two data streams from traffic detectors; namely, the *flow rate* and the *degree of saturation*. Throughout this paper, we refer to event detection when we analyse social media streams and we say anomaly or change detection when we analyse flow and saturation data.

Our method does two things. First we identify for every twitter event all road segments that could be directly or indirectly affected. Along those road segments we investigate the loop detectors for any changes in behaviour that could be attributed to the reported traffic issue. If there is a measurable impact on the joint expected transition probabilities in traffic flow and degree of saturation for a particular reported issue, we do expect this to be evidenced in a subset of the timeseries data. Usually, adjacent detectors are affected at the same or at similar time intervals, because up- and down-stream flows are correlated. Second, we find spatially correlated change intervals in the detector data. The change intervals (or anomolous regimes) are bounded by normal behaviour before the traffic incident and by the regime that returns to normal after the incident has fully cleared. Depending on the severity of the incident, anomolous events can be quite long.

Our change detection method can also be used without any traffic related tweet reported by authoritative sources. If any anomolous behaviour is detected in spatio-temporally correlated detectors, we can trigger an anomolous event that may need validation by the traffic operators. This way, changes can be reported statistically as they happen for which no social media evidence has been reported.

As an output of our method, the set of affected detectors are presented to the operator. Our contribution is a method that takes geo-coded tweets about urban traffic accidents and inductance loop timeseries data and produces a set of detectors for which we have evidence of an unusual event (accidents, etc.).

C. Analysis of our method

The objective of this analysis is to present spatially correlated anomalies to a traffic operator in an automatic fashion. We evaluate our approach using two datasets for each event that was identified and geo-coded using authoritative Twitter messages. One dataset includes the event time window and the validation dataset takes the same time of day and day of week for 3 weeks prior to the event into account. We show that we find significantly more events with higher severity scores in the event set compared to the validation set. However, this does not imply that the validation set is void of anomalies. For spatio-temporally correlated detectors that exhibit anomolous regimes, we query social media sources to find potential causes. Since we do not have the ground truth of all anomalies that happen across the entire city, we believe that these events shall still be presented to the operator in real-time for further investigation. The operator can discard these statistical anomalies, if no particular cause can be identified.

II. MODEL

In this section, we explain our notations. In order to motivate the choice of our anomaly detection method, we present a simplified model of flow and saturation signals (i.e., the sequences of detector measurements) as a Markov process. We also define the notion of *anomalous regime*.

A. Normal regime

Unless explicitly stated, we only consider a single detector. We model the evolution of detector measurements over time as a non-homogeneous Markov chain. The set S contains the possible values of detector measurements. The transition functions P_t at time t change over time¹. We assume that these functions $\{P_t\}$ are a priori unknown and we estimate them using historical data.

B. Anomalous Regimes

In this paper, we model anomolous regimes by intervals, $[\nu_{\text{start}}, \nu_{\text{end}}]$, which correspond to a start and an end point. The start, ν_{start} , and end, ν_{end} , of the interval are change points from and into normal regimes respectively.

We assume that there exist distinct processes Q_t and M_t , which we define later, such that $P_t = M_t$ if $t \notin (\nu_{\text{start}}, \nu_{\text{end}})$ and $P_t = Q_t$ if $t \in (\nu_{\text{start}}, \nu_{\text{end}})$. We say that two anomolous regimes, $[\nu_{\text{start},1}, \nu_{\text{end},1}]$ and $[\nu_{\text{start},2}, \nu_{\text{end},2}]$, are spatio-temporally correlated, if the detector locations are adjacent and the change intervals overlap in the time domain, i.e., $\cap_{i=1}^2 [\nu_{\text{start},i}, \nu_{\text{end},i}] \neq \emptyset$.

¹We consider only work week days (Monday to Friday), Saturday and Sunday.

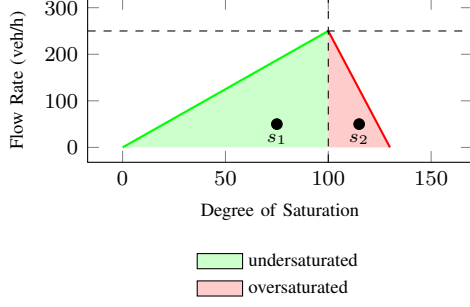


Fig. 1. Traffic Phase Diagram

III. OUR DATA

A. Inductance Loop Data

The system of over 2000 dedicated inductance loop sensor in Dublin is known as SCATS (Sydney-Coordinated Adaptive Traffic System). Each detector is associated with an intersection, for which we take the centroid as the geographic location. This allows us to issue a spatial query expressing that certain detectors are potentially affected by a traffic event for which we extracted location information, such as street names and neighbourhoods. Each sensor i generates the following sequences of measurements

$$\{s_t^i = (ds_t^i, f_t^i) : t = 1, 2, \dots\}.$$

The values ds_t^i, f_t^i can be interpreted as follows, with the help of the phase diagram for the loop detector data of Figure 1. This diagram gives the relationships between the two observable states of a detector. The x -axis corresponds to the degree of saturation, ds , which represents capacity. The y -axis gives the flow rate, f , in vehicles per hour. The diagram mainly shows two regimes of the evolving traffic state. As the degree of saturation increases, the flow rate increases up to a maximum value at a degree of saturation of 100. As the degree of saturation increases further, we reach an oversaturated regime with generally a sharp dropoff in the flow rate. A state tuple is emitted after a cycle is completed and the detector data are aggregated over all phases. Each phase in a cycle gives pre-configured paths across the intersection the opportunity to service vehicles (see Figure 2).

If the traffic state is currently observed as undersaturated, then the supply of vehicles from upstream to that intersection limits the number that can be serviced (state s_1^i). While in the oversaturated regime (red region in figure 1), the demand is the limiting factor due to spill-over effects from downstream (state s_2^i). This may occur during accidents when congestion back propagates and under-saturated conditions propagate downstream.

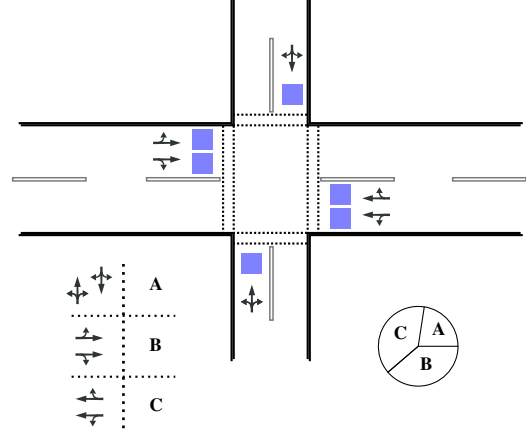


Fig. 2. Intersection Phase Diagram

B. Social Media Data

We use a dataset composed of Twitter messages from three reputable sources of traffic information in Dublin: LiveDrive, AA Roadwatch and Garda Traffic, over the months of January to July 2013. These Twitter messages are used to curate a dataset that is taken as the ground truth of traffic related events such as concerts, collisions and protests. These events are labelled according to the keywords in the Twitter message.

A GPS coordinate is manually assigned to each Twitter message by a human expert, thereby minimizing errors. An example message is “LiveDrive: Emergency services are attending an incident on Westmoreland St after the junction of Fleet St. The two right lanes.” received on the 25th day of April 2013 at 16:14 hours, which is assigned the GPS coordinates (53.34, -6.26).

Formally, a tweet is given as an array

$$W = \langle t, loc, type, description \rangle, \quad (1)$$

where t is the time stamp of the tweet, loc is the location given in longitude and latitude, the type indicates whether the tweet raises an “alert” message, and finally a description string which is not used in our analysis.

This curation can be generalised using the approach described in [13], where a dictionary is used to classify the content of the message into different event categories, which we leave for future work.

IV. ANOMALOUS REGIME DETECTION

Our method relies on both statistical change-detection and additional information from experts through Twitter. We employ an algorithm similar to CUSUM adapted to Markov chains (cf. [14]). The original CUSUM algorithm for piecewise-i.i.d. sample sequences is not suitable for road traffic measurements

because human transportation patterns are inherently time-varying. First, we consider the case where there are two possible regimes. Let $M_t(y | x)$ denote the transition probability in the normal regime. Let $T_w(s_t)$ denote the transition probability of the current regime over the past w transitions, i.e., considering all states $s \in \{s_{t-w}, \dots, s_t\}$. Let $knn(s; t-2)$ denote the set of k nearest neighbours of s among $\{s_1, \dots, s_{t-2}\}$, and let the corresponding set of indices be $\Phi_{t-1}(s) = \{j : s_j \in knn(s; t-2)\}$. In order to estimate these probabilities, we use the k -nearest neighbour transitions to estimate the transition probability in the normal regime and we estimate the transition probability in the current regime over the recent w observed state transitions as follows:

$$\widehat{M}(s | s_t) = \frac{1}{k} \sum_{\ell \in \Phi_{t-1}(s_t)} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(s_{\ell+1}-s)^2}{2\sigma^2}} \quad (2)$$

$$\widehat{T}_w(s_t) = \frac{1}{w} \sum_{\ell=t-w}^{t-1} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(s_{\ell+1}-s_\ell)^2}{2\sigma^2}}, \quad (3)$$

where the sum is over next-states $s_{\ell+1}$. The size of window parameter, w , determines the sensitivity of the anomaly detection algorithm. As the window size gets larger, more and more past information is incorporated into the estimation of the current regime transition probability, which averages out potentially short-lived anomalies.

It is well-known that kernel density estimation is consistent [15], i.e., the estimates converge to the true distributions.

We consider the model where the state space S is an interval in the entire real line \mathbb{R} . Our approach is similar to [2]. Generally, change-detection algorithms have a threshold parameter that must be set. This has practical implications of the algorithm as the threshold value may need to be tuned a priori for each detector or type of event we are aiming to detect. A further assumption is that the underlying process being modelled is stationary. This assumption, however, does not hold in traffic because of network effects and other variables that are difficult to model, such as weather. Moreover, the dynamics of traffic change from year to year. For example in Dublin we have observed an increase in flow and degree of saturation over the last 3 years. As a consequence, we do not employ a threshold rule like in CUSUM algorithms. Instead we relax the assumption of stationarity and identify regime changes in the cumulative sum (equation 5) of the log-likelihood ratios (equation 4).

$$R_t = \log \left[\frac{\widehat{T}_w(s_t)}{\widehat{M}(s | s_t)} \right] \quad (4)$$

$$S_t = S_{t-1} + R_t. \quad (5)$$

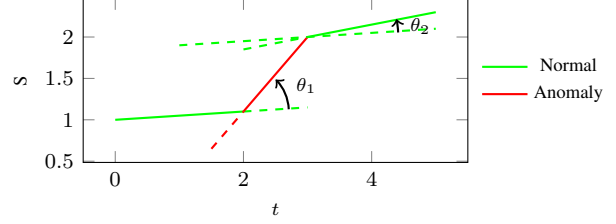


Fig. 3. Anomalous Regime Detection. An anomalous regime is bounded by a normal regime before the event occurred and a normal regime after the anomalous traffic pattern returns to normal.

Given the cumulative sum of the log-likelihood ratios, we fit piece-wise linear regression curves, $\widehat{S} = \alpha + \beta S$, and their corresponding 95% prediction intervals, $w(t)$, over the time domain $[\tau, t]$, where τ is the last found change point. A curve is bisected given the following rule

$$\arg \max_{\tau' \in [\tau, t]} b(\tau') := \left\{ \left| (\widehat{S} - S_{\tau'}) \right| > w(\tau') \right\}. \quad (6)$$

A change point that bisects the regression curve is found when there are residuals that exceed the prediction interval. The largest residual is chosen as the new change point, τ' . In this case a regime is fitted over the domain $[\tau, \tau']$ and stored. Both end points mark changes to the prior and the successive regimes. We say a regime is left-bounded until a new change point is found, which is the case for $[\tau', t]$. This interval is also the real-time regime, as the time t increments with every new measurement until a new bisection point is found.

Importantly, these change points may only account for minor behavioural changes compared to a model. In order to evaluate the magnitude in a principled way, we look at the angle, θ between two successive behavioural regimes. Since we define each regime to be a linear regression function, $\widehat{S} = \alpha + \beta S$, we can use the slopes to calculate θ . Note, that the angle satisfies the inequalities $0 \leq \theta \leq \pi/2$, which allows us to define a threshold value, δ , in order to detect anomalous regimes.

Once an anomalous start of a regime is detected, we observe the anomaly until it returns to normal behaviour. We define the return to normal behaviour with respect to the behaviour of the piece-wise linear regression curve before the anomalous regime started. Figure 3 schematically depicts the rule of identifying the entire anomalous regime given two threshold values, δ_1 and δ_2 , where $\delta_1 > \delta_2$. If the angle between the current and previous regime exceeds δ_1 , then we enter an anomalous regime. We stay in this anomalous regime until the angle between the normal regime and the current regime is smaller than δ_2 .

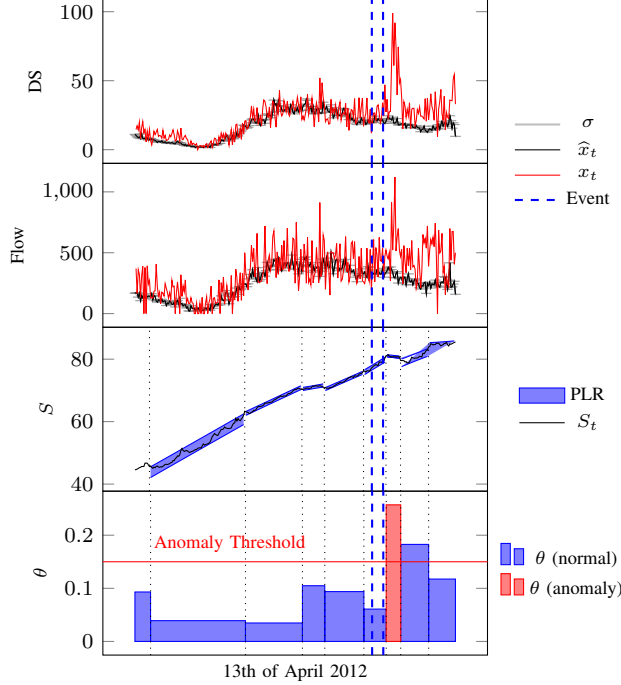


Fig. 4. Timeseries and corresponding change vector for a single detector on a day of an event. The two top rows show the degree of saturation and flow of a given detector. The two bottom rows show the change vector with piece-wise linear regimes and the angle, θ , between those regimes respectively.

Figure 4 shows the result of our methodology on a particular detector of a given day. The first two rows show the actual flow and degree of saturation (red curves) compared to a historic model (black curves). The vertical dashed blue lines represent the start and end of the event given the corresponding Twitter messages. The third row shows the cumulative sum of the log-likelihood ratios and their piece-wise linear regression lines and the corresponding angles between them in the last row of the figure. In this particular case, the anomalous regime occurs just after the event is marked as cleared, which is evidenced by the angle, θ , exceeding a defined anomaly threshold. The regime that immediately follows the anomalous one exhibits a similar angle (i.e., $\theta < \delta_2$) compared to the regime just before the anomalous regime. Consequently, we can indicate to the operator not only the severity of the anomalous regime given the angle θ , but also when a regime returns to normal behaviour.

Additional to individual detector information, the spatial correlation of these anomalies carries signals that can be exploited to understand the spread of anomalous traffic conditions. For example, an accident inevitably impacts at least two detectors. Upstream detectors experience and increased density due to congestion building

up backwards, while downstream detectors receive less flow than expected. As a consequence, the degree of saturation drops alongside flow.

V. EVALUATION

A. Preliminaries

This section serves to validate the anomalies we have found using our approach. Generally, the evaluation of heterogeneous data with drifts in the timeseries is challenging and a dichotomous answer to false positives and false negatives is not easily or even impossible to come by. This is due to the partial observability of traffic. Ground truth information can only be considered if road segments are surveyed over the evaluation period.

Because of these challenges, we follow a strategy to validate anomalies as much as possible through external sources. These include Twitter, road works and event reports. Using Twitter we draw upon messages from the general public and authoritative sources. As mentioned before authoritative sources present a semi-structured format that includes geographic references that can be used to identify the adjoining streets that are potentially affected. For events, floods, and road works, messages by the general public can also be useful.

B. Experimental Setup

In total we evaluated 33 events that may or may not have a measurable impact on traffic in the immediate vicinity of the event location. Events can include accidents as reported from authoritative sources, flooding events, concerts and sports matches. All events tend to be short-term events of a couple of minutes to a couple of hours. Concerts and matches are planned events, where road closures and diversions may be partly employed. From an operational perspective, these events do not necessarily amount to unknown anomalies, because they are planned weeks in advance. Accidents and floods occur spontaneously. As a consequence the operator is interested in detecting the emergence of anomalies so that mitigation strategies can be implemented before entire neighbourhood are at risk of becoming congested.

Additional to the event dataset, we also prepared a validation set. For each event we compiled a validation dataset consisting of the same weekday and time of day for the three weeks prior to the event. We use the validation set to test the anomaly detection algorithm without any prior confirmation of traffic related Twitter messages. For any anomalies we find, we turn to the Twitter messages to find any potential causes that have been published.

We investigated events that started between 7am and 7pm and ran our anomalous event detection algorithm

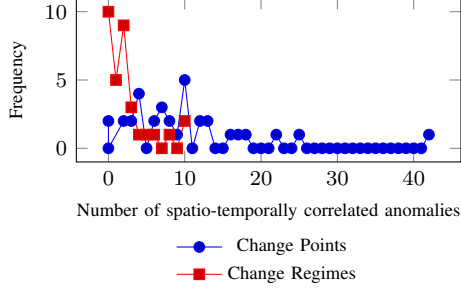


Fig. 5. Number of change regimes and change points and their frequencies across the dataset.

starting from midnight the day before the event until midnight of the day of the event. However, for change points or anomalous regimes to be assigned to an event, the time at which the anomaly was detected has to fall into the event time frame, which we define as 30 minutes before the event was reported to have started and 30 minutes after the event has been reported to have finished. We set the window size, w , to calculate the current regime transition probability to be 10, which amounts to the last 50 minutes of the timeseries of a detector. A detector emits a new state every 5 minutes. The transition probabilities of normal regimes are trained on a year's worth of historic data prior to the event with k of the k -nearest neighbour algorithm set to 100. The anomaly threshold for the angle, δ_1 , is set to 0.15 and the angle threshold for which anomalous events return to normal behaviour is set to 0.1.

Since, we do not have ground truth information about anomalies, we chose the parameters for the algorithm through experimentation. We selected a few events for which we knew that there was a significant impact on traffic and manually tuned the parameters so that these anomalies are found. With ground truth information, however, we could optimise the choice of parameters in order to minimise false positives and false negatives.

C. Results

We first consider an end of day analysis, which means we search in the historic change detection vectors, i.e., the cumulative sum (equation 5) of the log-likelihood ratios (equation 4), for change anomalies with a distinct point of when the anomaly starts and another change point when this anomaly returns back to normal behaviour (see red curve in figure 5). For 10 events we found no anomalous regimes as well as no change points (blue curve in figure 5) within the event time frame. Investigating the tweet messages, however, does reveal that some impact on traffic has been reported. The reason, that the detection algorithm has neither confirmed an anomalous regime nor change points can

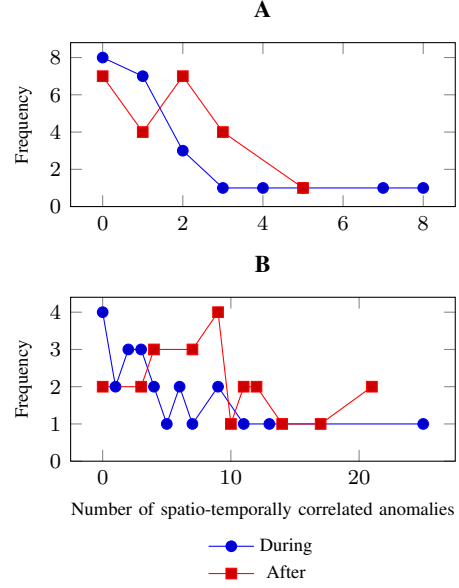


Fig. 6. Number of change regimes (A) and change points (B) detected during and after the event time frame.

be due to characteristics external to the model, such as weather. If heavy rain causes anomalies, they are likely to be evidenced before the event was recorded. Otherwise, the impact may go unnoticed due to the anomaly threshold.

For the remaining 23 events that had anomalous regimes, eight events exhibited anomalous behaviour after the event time frame (see figure 6 A). Seven events had only 1 detector for which an anomalous regime started during the event time frame. This implies that similar anomalies are not observed in up- or downstream detectors. However, change points occur more frequently during event time frames (see figure 6 B).

After evaluating the workflow of first identifying an event through authoritative Twitter messages and then looking for their impact, we now follow the reverse approach. We run our change detection algorithm over the detector timeseries of this validation set and found 201 anomalous regimes. We found evidence of some event for 94 (out of 201 possible candidates) reported on Twitter and road closure reports issued by the Dublin city council.

Figure 7 provides the number of events grouped by types. Two road works coincided with a concert event. Almost half of the anomalous regimes found are spatio-temporally correlated. I.e., they overlap in time and are spatially adjacent (see figure 8).

Interestingly, we found events where anomalies are evidenced before the event started and after it finished.

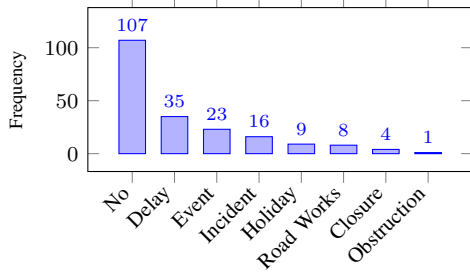


Fig. 7. Number of anomalous regimes per event type of the validation set

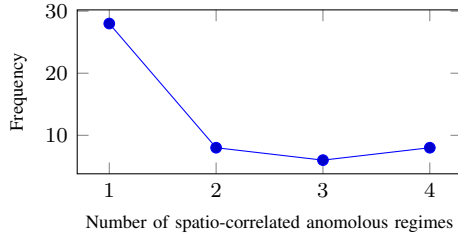


Fig. 8. Frequency of the number of spatio-correlated anomalous regimes in the validation dataset

One example is an exam taking place for several hundred students. While the exam was in progress, traffic returned to normal, which was correctly identified by our algorithm. Other events, such as national soccer games, tend to show anomalies for the entire duration of the event due to traffic restrictions.

The identified related Tweet reads as follows:
 N11: Remember that exams take place in the RDS this morning so expect higher volumes of traffic than usual
 Source: LiveDrive
 Date: 08/05/2013 07:46:15

VI. CONCLUSION

User generated content is extremely noisy making it similarly challenging to identify validated, reliable event related content. Additionally, events identified through the social media stream may in fact have little or no impact on traffic volumes. For example, a collision may be reported and the same collision may have been cleared minutes after the initial alert causing no visible congestion.

We presented a change detection algorithm for urban traffic anomalies, which uses traffic alert messages sent by traffic authorities to define a spatial scope of potential impacts in the vicinity of the accident. In 75% of the alert messages, anomalies can be found in related detectors and thus operators can assess the impact of the reported accident. This provides operators a tool

to automatically verify and monitor reported accidents and take action as soon as impacts can be confirmed. What is otherwise a significant manual effort, can now be computed automatically.

We also tested our algorithm on a validation set with no prior social media messages reporting an accident. In almost 50% of the detected anomalies we were able to confirm traffic incidents according to social media reports. Hence we showed, that our detection method can be used in real-time to monitor for traffic anomalies for all detectors at the same time. As anomalies emerge, the operator can be informed for further investigations.

By combining these two noisy information streams, anomalous traffic events may be identified with greater confidence and also assigned a human readable label to be displayed to traffic monitors.

REFERENCES

- [1] K. Sricharan, R. Raich, and A. O. Hero, "k-nearest neighbor estimation of entropies with confidence," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. IEEE, Jul. 2011, pp. 1205–1209. [Online]. Available: <http://dx.doi.org/10.1109/isit.2011.6033726>
- [2] T. L. Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *Information Theory, IEEE Transactions on*, vol. 44, no. 7, pp. 2917–2929, Nov. 1998.
- [3] V. Guralnik and J. Srivastava, "Event detection from time series data," in *KDD*, 1999, pp. 33–42.
- [4] J. Lin, M. Vlachos, E. J. Keogh, and D. Gunopulos, "Iterative incremental clustering of time series," in *EDBT*, 2004, pp. 106–122.
- [5] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," *Transportation Research Part B*, vol. 18, pp. 1–11, 1984.
- [6] F. Schnitzler, T. Liebig, S. Mannor, and K. Morik, "Combining a gauss-markov model and gaussian process for traffic prediction in dublin city center," in *Proceedings of the Workshops of the EDBT/ICDT 2014 Joint Conference*, 2014.
- [7] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg, "Chatter on the red: what hazards threat reveals about the social life of microblogged information," in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, ser. CSCW '10. New York, NY, USA: ACM, 2010, pp. 241–250.
- [8] H. Becker, M. Naaman, and L. Gravano, "Beyond Trending Topics: Real-World Event Identification on Twitter."
- [9] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 851–860.
- [10] A. M. Popescu, M. Pennacchiotti, and D. Paranjpe, "Extracting events and event descriptions from Twitter," in *Proceedings of the 20th international conference companion on World wide web*, ser. WWW '11. New York, NY, USA: ACM, 2011, pp. 105–106.
- [11] J. Nichols, J. Mahmud, and C. Drews, "Summarizing sporting events using twitter," in *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, ser. IUI '12. New York, NY, USA: ACM, 2012, pp. 189–198.

- [12] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence, "Improving traffic prediction with tweet semantics," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, ser. IJCAI'13. AAAI Press, 2013, pp. 1387–1393.
- [13] E. M. Daly, F. Lécué, and V. Bicer, "Westland row why so slow?: fusing social media and linked data sources for understanding real-time traffic conditions," in *IUI*, 2013, pp. 203–212.
- [14] L. Gereneser and G. Molnar-Saska, "Change detection of hidden markov models," in *Decision and Control, 2004. CDC. 43rd IEEE Conference on*, vol. 2. IEEE, Dec. 2004, pp. 1754–1758 Vol.2.
- [15] E. Giné and A. Guillou, "Rates of strong uniform consistency for multivariate kernel density estimators," *AMS*, 2001.